Natural Language Processing Final Project Report

Tweet Hashtag Recommendation

Chuhan Liu

1. ABSTRACT

This project mainly works on tweets hashtags recommendation. When twitter users want to post tweets, this project can generate several relevant hashtags for users to choose and add to their tweets. In this project, I use a model inspired by classic TF-IDF model to get the most 10 relevant hashtags to a tweet which has not been tagged with hashtag yet. In contrast to some popular approaches such as Naïve Bayes model, which based on all words and sequence of words in a tweet, this method mainly focuses on the correlation between a particular word in the tweet and the hashtag corresponding to the tweet.

2. INTRODUCTION

2.1 Background

Twitter is one of the most popular micro – blogging sites in the world. Millions of people express their opinions on many particular topics, show interesting things in their life and follow up hot topics all around the world. However different to other SNS, like Facebook, Instagram, etc. which mainly provide platforms for people to show routines in their life and know more about their friends on the internet, Twitter is also a kind of very important resource for people to get information and know what is going on in the world. Then the hashtags of tweets which can group the tweets with same topics together makes it very easy for people to search the current hot topics and get news. However, not every twitter user tags a hashtag to his/her tweet when he/she posts it. According to the statistics, only %8 tweets contain a hashtag.[1] So, assigning hashtags to those tweets without hashtags is very useful and more convenient for Twitter users.

2.2 Related Work

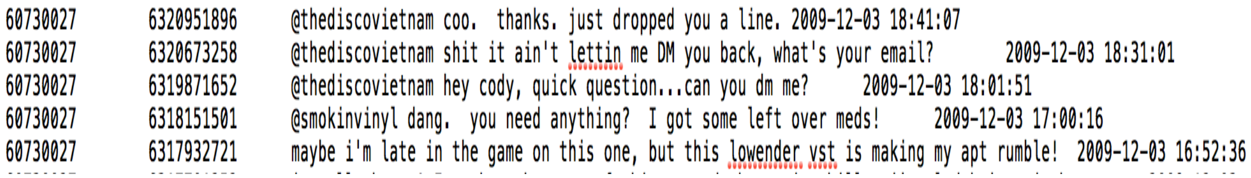
A lot of works have been done on hashtags suggestion for tweets. A very intuitive approach to recommend hashtags for tweets is using TF-IDF model and representing each tweet in word vector space. Then when we get a new tweet without hashtag, we can compare the similarity of the vector corresponding to this tweet with all vectors in the TF-IDF matrix and use the hashtag of the tweet, which is already tagged with a hashtag and most similar to the new tweet, as the hashtag for the new tweet.[2] Naïve Bayes rule is another very common method to solve this problem. Mazzia and Juett computed the conditional probability of the tweet based on the sequence of the words, and use maximum a posteriori (MAP) as the rule to classify tweets into different hashtags classes.[3] Comparing WordNet similarity between words in similar tweets has also been introduced to the field of hashtag recommendation by Li and Wu.[4] This model has the similar idea to the model used in this project, which are all based on the singular word in each tweet. Most of approach used to hashtag recommendation is a kind of supervised learning, but there are also few approaches based on unsupervised learning. Fréderic Godin and Viktor Slavkovikj used topics models, LDA and Gibbs sampling to assign hashtags for the tweets on which have been implemented POS.[5]

3. DATASET AND DATA PRE-PROCESS

3.1 DATASET

The dataset was collected by Z. Cheng, J. Caverlee, and K. Lee[6] from September 2009 to January 2010 which contains training set data and test set data. In this project, I only use the training set data in this project which contains 115,886 Twitter users and 3,844,612 updates from the users. All the locations of the users are self-labeled in United States in city-level granularity.

The raw dataset is like:



```
60730027	6320951896	@thediscovietnam coo.  thanks. just dropped you a line. 2009-12-03 18:41:07
60730027	6320673258	@thediscovietnam shit it ain't lettin me DM you back, what's your email?	2009-12-03 18:31:01
60730027	6319871652	@thediscovietnam hey cody, quick question...can you dm me?	2009-12-03 18:01:51
60730027	6318151501	@smokinvinyl dang.  you need anything?  I got some left over meds!	2009-12-03 17:00:16
60730027	6317932721	maybe i'm late in the game on this one, but this lowender vst is making my apt rumble!  2009-12-03 16:52:36
```

Each tweet is in one line. And there are 4 elements separated by '\t' for each tweet. The first string of number is the user ID; the second string of number is the tweet ID; the third element is the content of the tweet; the last element is the time when the tweet is posted. In this project, I only use the third element content of tweet for hashtag recommendation.

3.2 Data Pre-Process

Before I start to train the tweets dataset, I firstly implement some necessary data wrangling steps on the raw tweet contents. (1) make all letters in the tweet content to lowercase (2) extract all distinct hashtags in each tweet (3) remove all punctuations from tweets excluding "#" and "@" (4) remove all "@other users" phrases in tweets (5) remove all digits, non-English, non-ascii characters in tweets (6) look for those "#words" phrases in the middle of a tweet and change them to regular word without "#", remove other "#words" phrases at the end of tweets (7) expand stopwords by adding some words like "rt", preposition, modal particle like "OMG", "Wow", "Wa', etc. , and remove all stopwords in the tweet. (8) stem the remain word using Porter Stemmer in Python NLTK library

3.3 Training Data and Test Data

After data pre-processing, there are total 405,511 tweets having hashtags out of the initial 3,844,612 tweets. Among these 405,511 tweets, I randomly extract 200 tweets from it, delete the hashtags tagged with these tweets and use these 200 tweets as the test data. Then the remaining 405,311 tweets are training data.

4. HASHTAG RECOMMENDATION IN CLASSIC TF-IDF MODEL

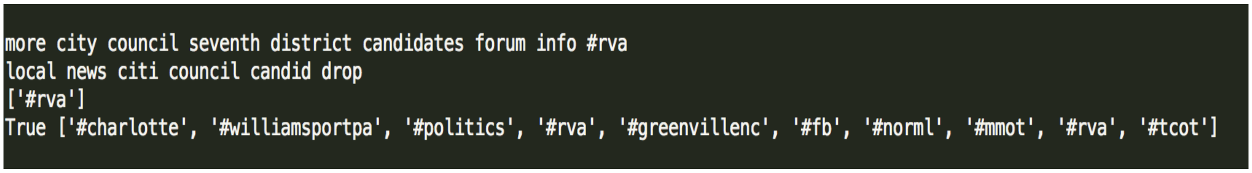
4.1 Implement Process

In this section, I represent all tweets in the training data in word vector space and calculate the training TF-IDF matrix for these tweets. Similar to this work on training data, I also transform each tweet in the test data into a TF-IDF vector. Then for each test tweet TF-IDF vector, I compute the cosine similarity between the test tweet vector and all training tweets vector in the training TF-IDF matrix. Finally, I will extract ten hashtags tagged with the several top training tweets which are most similar to the test tweet.

For TF-IDF model, I use the default parameter of the max document frequency and minimum document frequency, and just implement 1-gram to extract the word features. So, finally, I totally get 158,270 word features.

4.2 Result

One of the test tweet I extract randomly from the initial dataset as the test tweet is like: *"rt @thetvhangman #ff @spiraluptoday @missyklok @localmohaw @xtexasbelle @c0rpsebunny @glaminmotion"*, it obviously cannot be recommended hashtags based on a tweet without actual body content, after data wrangling, this tweet becomes empty. So now, I have 199 test tweets. Among these 199 tweets, I make correct hashtags recommendation for 103 tweets. The output is like:



```
more city council seventh district candidates forum info #rva
local news citi council candid drop
['#rva']
True ['#charlotte', '#williamsportpa', '#politics', '#rva', '#greenvillenc', '#fb', '#norml', '#mmot', '#rva', '#tcot']
```

The first line is the original test tweet with its hashtag; the second line is the most similar tweet to the test tweet after data wrangling in training dataset; the third line is the original hashtags of the test tweet; the fourth line is the 10 hashtags recommended to user.

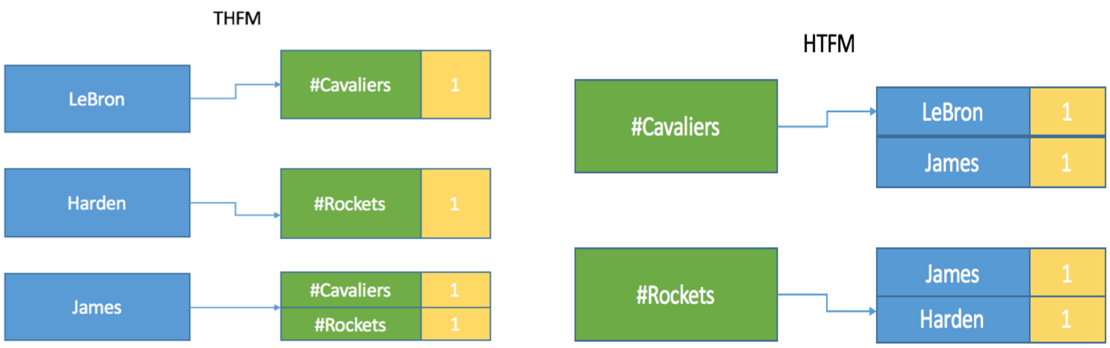
In the following sections, I will discuss about why for some test tweets it’s hard to give the same hashtags as it originally has.

5. HASHTAG RECOMMENDATION IN HF-ICF MODEL

5.1 Hashtag Frequency and Inverse Corpus Frequency Model

According to the correlation between tweets contents and their hashtags, for most of time, we can find out that there are always some particular words in tweets which narrow down the scope of the theme of the tweets. For example, a tweet like “LeBron showing his respect and love for Isaiah Thomas. #BiggerThanBasketball #GameRecognizeGame #Celtics #Cavs”[7] , even though the words like Celtics and Cavaliers don’t appear in the body of tweet, it is obviously to see that this tweet is about NBA, Celtics and Cavaliers. So for hashtag recommendation, instead of finding the similarity between different tweets, it will also be a good approach if we can find the relations between particular words and hashtags. Then in this section, a model inspired by classic TF-IDF model but based on singular word will be implement to try to solve the problem of hashtag recommendation.[8]

Firstly, I need to build two maps showing the statistical counts relations between words and hashtags. One I call it Term Hashtag Frequency Map and the other one I call it Hashtag Term Frequency Map.



THFM: Use distinct words in tweets as keys, and map all distinct hashtags and their appearing time when the word is in the tweet body while the hashtag is tagged with this tweet.

HFTM: Use distinct hashtags tagged with tweets as keys, and map all distinct words and their appearing times when the hashtag is tagged with the tweet while the word is in this tweet

5.1.1 Hashtag Frequency

Hashtag Frequency analogous to Term Frequency:

$$hf_{t,h} = \frac{THFM[t][h]}{\sum_{h'} THFM[t][h']}$$

The hashtag frequency refers to the frequency of a hashtag when a particular word appears in a tweet. So for each word(term), the hashtag frequency for a particular hashtag is the ratio of the times of the hashtag appears when the word is in the tweet to the total number of all hashtags appear when the word is in tweets.

5.1.2 Inverse Corpus Frequency

Inverse Corpus Frequency analogous to Inverse Document Frequency:

$$icf_h = \log \frac{|Corpus|}{\sum_t HTFM[h][t]}$$

The inverse corpus frequency aims to discount the weight of the hashtags which appear many times in the corpus. For a particular hashtag, the fraction part in logarithm is the ratio of the size of the corpus to the total numbers of words mapped with this hashtag.

5.1.3 Score of Word and Rank Rule

Using the training data, I can make two dictionaries for computing the score for each word when I get a new input tweet. One dictionary stores the hashtag frequency value for different hashtags mapped with different words. The other dictionary stores the inverse corpus frequency values for all hashtags. Then the score for a word in the input tweet will be:

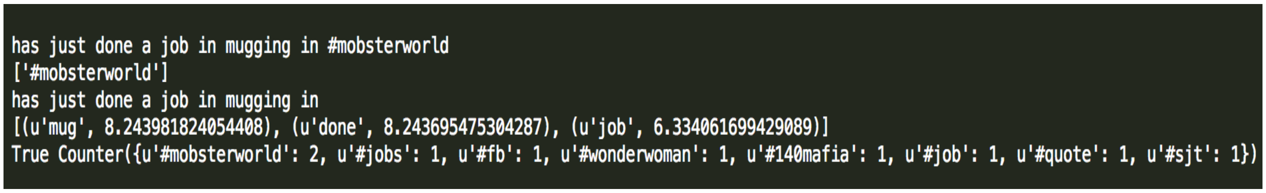
$$S_{w_i} = \sum_{h} hf_{w_i,h} \times icf_h$$

For each word, I compute the sum of the product of hashtag frequency and the inverse corpus frequency of every hashtag mapped with this word as the score of this word. If I cannot find this word in our training data, I set the score of this unseen word to 0. So, after assigning scores to all the remain words of the pre-processed input tweet, I can assume that there may be a stronger correlation between the word and the theme of the tweet, if the word gets a higher score. After I get the few top words with highest scores, I extract the top hashtags with the highest product value of hashtag frequency and inverse corpus frequency as the recommended hashtags. I still extract 10 hashtags as the recommended hashtags for a new tweet.

5.2 Result

As the same reason I have mentioned in the previous 4.2 section, there are actually 199 tweets for my 200 random test tweets sample. There are total 92 tweets whose original hashtags are in the scope of my recommended hashtags. If I compute the accuracy as what we commonly do for supervised learning, the accuracy will be 0.46. It seems to be a very poor prediction in the supervised learning. However, in the case of hashtag recommendation, we cannot treat it as the common supervised learning. I will discuss the reason in the following section.

The output for each tweet is like:

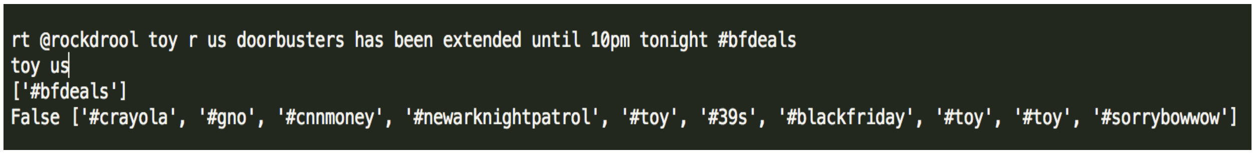


```
has just done a job in mugging in #mobsterworld
['#mobsterworld']
has just done a job in mugging in
[(u'mug', 8.243981824054408), (u'done', 8.243695475304287), (u'job', 6.334061699429089)]
True Counter({u'#mobsterworld': 2, u'#jobs': 1, u'#fb': 1, u'#wonderwoman': 1, u'#140mafia': 1, u'#job': 1, u'#quote': 1, u'#sjt': 1})
```

The first line is the original tweet; the second line is the original hashtags tagged with tis tweet; the third line is the tweet after pre-processing; the fourth line is the words ranked with their scores; the last line is the 10 hashtags I recommend to this tweet in a Counter Collection format, the larger the counts value is, the more highly that hashtag is recommended.

6. CONCLUSION

The process of hashtag recommendation in this project seems to be a supervised learning process. I use tweets with hashtags as the training data, and assign hashtag for those new tweets based on this training set. However, in contrast to the rule we use to judge a supervised learning model, it is not fair to use the accuracy of correct prediction in hashtag recommendation. Since, tweet is a kind of open context sentence. Based on a particular hashtag, Twitter users can make comment on this topic or just express their own emotions. Sometimes, it is really hard to decide the theme or narrow the scope of the tweet by just analyzing the body of the tweet. On the other hand, a particular tweet can also involve many information, but the Twitter user can just tag it with a particular topic. For example:

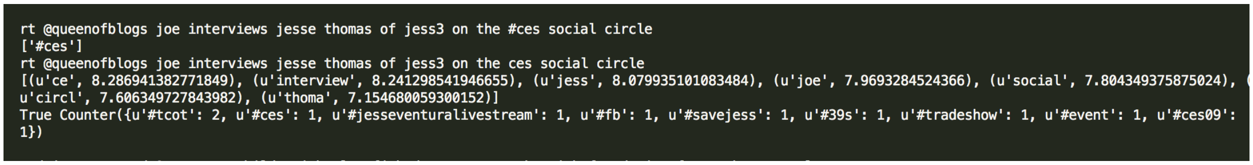


```
rt @rockdrool toy r us doorbusters has been extended until 10pm tonight #bfdeals
toy us
['#bfdeals']
False ['#crayola', '#gno', '#cnnmoney', '#newarknightpatrol', '#toy', '#39s', '#blackfriday', '#toy', '#toy', '#sorrybowwow']
```

In this example, my model gives the wrong recommendation compared with the original hashtag. However, it also makes sense if the user tags “#toy”, “#blackfriday” after the tweet. For a same tweet, like this tweet is about sale, toy and shopping, there are still various ways to tag it.

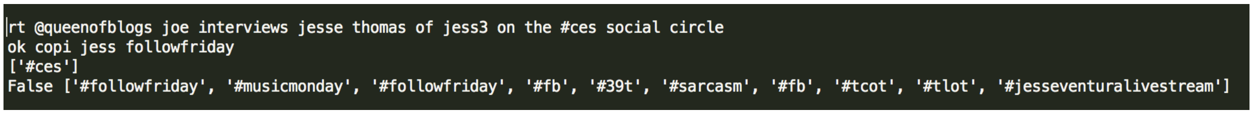
As to the two models TF-IDF and HF-ICF I use to recommend the hashtags, from the result we can see that TF-IDF model is better than the HF-ICF model. I think this indeed follows our intuitions. Although I have imagined that a particular word in the tweet may narrow the scope of the theme of the tweet, it is also very common that there is no proper name or some other words not very commonly appear in daily conversation mentioned in tweets. In this case, if we just extract the word from its context and try to find the correlation between the word and hashtags, it may just give us hashtag with most hashtag frequency. However, for some particular cases the HF-ICF model may perform better than the TF-IDF does. For example:

HF-ICF:



```
rt @queenofblogs joe interviews jesse thomas of jess3 on the #ces social circle
['#ces']
rt @queenofblogs joe interviews jesse thomas of jess3 on the ces social circle
[(u'ce', 8.286941382771849), (u'interview', 8.241298541946655), (u'jess', 8.079935101083484), (u'joe', 7.9693284524366), (u'social', 7.804349375875024), (
u'circl', 7.606349727843982), (u'thoma', 7.154680059300152)]
True Counter({u'#tcot': 2, u'#ces': 1, u'#jesseventuralivestream': 1, u'#fb': 1, u'#savejess': 1, u'#39s': 1, u'#tradeshow': 1, u'#event': 1, u'#ces09':
1})
```

TF-IDF:



```
rt @queenofblogs joe interviews jesse thomas of jess3 on the #ces social circle
ok copi jess followfriday
['#ces']
False ['#followfriday', '#musicmonday', '#followfriday', '#fb', '#39t', '#sarcasm', '#fb', '#tcot', '#tlot', '#jesseventuralivestream']
```

From those two results, we can see that TF-IDF model pays more attention on the similarity words between two tweets. So the hashtags TF-IDF model recommends are more about “jess” and “followfriday”. In contrast, HF-ICF pays more attention on each word in the tweet, thus, it can find the ‘ces’ in the tweet and assign high weight to this very uncommon word.

Thus, I can conclude that TF-IDF model does better job when words in tweets are tightly connected in the context, while HF-ICF performs better when the theme of the tweet is strongly related to one or more particular words in the tweet.

7. REFERENCE

[1]. S. M. Kywe, T.-A. Hoang, E.-P. Lim, and F. Zhu, "On Recommending Hashtags in Twitter Networks," in The 4th Int. Conference on Social Informatics, 2012.

[2]. E. Zangerle, W. Gassler, and G. Specht, "Recommending#-tags in twitter," in Proceedings of the Workshop on Semantic Adaptive Social Web, 2011.

[3]. A. Mazzia and J. Juett, "Suggesting Hashtags on Twitter," tech. rep., Computer Science and Engineering, University of Michigan, 2009.

[4]. T. Li, Y. Wu, and Y. Zhang, "Twitter hash tag prediction algorithm," in ICOMP'11 - The 2011 International Conference on Internet Computing, 2011.

[5]. Fréderic Godin, Viktor Slavkovikj, Wesley De Neve, Benjamin Schrauwen and Rik Van de Walle, "Using Topic Models for Twitter Hashtag Recommendation", WWW '13 Companion Proceedings of the 22nd International Conference on World Wide Web, 2013

[6]. Z. Cheng, J. Caverlee, and K. Lee, "You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users", 19th ACM Conference on Information and Knowledge Management (CIKM), 2010

[7]. Cited tweet from https://twitter.com/hashtag/GameRecognizeGame?src=hash

[8]. Eriko Otsuka, Scott A. Wallace and David Chiu, "A hashtag recommendation system for twitter data streams", Computational Social Networks, 2016